Report on the TREC 2004 Genomics Track

William R. Hersh
Department of Medical Informatics &
Clinical Epidemiology
Oregon Health & Science University,
Portland, OR, USA
hersh@ohsu.edu

Note: The TREC Genomics Track has a new URL, *http://ir.ohsu.edu/genomics/*. A more complete overview of the TREC 2004 Genomics Track (Hersh, Bhuptiraju et al., 2004) is available at this site as well as the main TREC site, *http://trec.nist.gov/*.

1 Background

The goal of the TREC Genomics Track is to create test collections for evaluation of information retrieval (IR) and related tasks in the genomics domain. The Genomics Track differs from the other TREC tracks in that it is focused on retrieval in a specific domain. Initially the track has focused on modeling advanced users accessing the scientific literature. The advanced users include biomedical scientists and database curators or annotators. New advances in biotechnologies have changed the face of biological research, particularly "high-throughput" techniques such as gene microarrays (Mobasheri, Airley et al., 2004). These not only generate massive amounts of data but also have led to an explosion of new scientific knowledge. As a result, this domain is ripe for improved information access and management.

The scientific literature plays a key role in the growth of biomedical research data and knowledge. Experiments identify new genes, diseases, and other biological processes that require further investigation. Furthermore, the literature itself becomes a source of "experiments" as researchers turn to it to search for knowledge that drives new hypotheses and research. Thus there are considerable challenges not only for better IR systems, but also for improvements in related techniques, such as information extraction and text mining (Hirschman, Park et al., 2002).

Because of the growing size and complexity of the biomedical literature, there is increasing effort devoted to structuring knowledge in databases. The use of these databases is made pervasive by the growth of the Internet and Web as well as a commitment of the research community to put as much data as possible into the public domain. One of the many key efforts is to annotate the function of genes. To facilitate this, the research community has come together to develop the Gene Ontology (GO, www.geneontology.org) (Anonymous, 2004). A major use of the GO has been to annotate the genomes of organisms used in biological research. The annotations are often linked to other information, such as literature, the gene sequence, the structure of the resulting protein, etc.. An increasingly common approach is to develop "model organism databases" that bring together all this information in an easy to use format. Some of the better known model organism databases include those devoted to the mouse (Mouse Genome Informatics, MGI, www.informatics.jax.org) and the yeast (Saccharomyces Genome Database, SGD,

www.yeastgenome.org). These databases require extensive human effort for annotation or curation, which is usually done by PhD-level researchers. These curators could be aided substantially by high-quality information tools, including IR systems.

The TREC 2004 Genomics Track consisted of two tasks. The first task was a standard ad hoc retrieval task using topics obtained from real biomedical research scientists and documents from a large subset of the MEDLINE bibliographic database. The second task focused on categorization of full-text documents, simulating the task of curators of the Mouse Genome Informatics (MGI) system and consisting of three subtasks. One subtask focused on the triage of articles likely to have experimental evidence warranting the assignment of GO terms, while the other two subtasks focused on the assignment of the three top-level GO categories. The track had 33 participating groups, the most of any track in TREC 2004.

A total of 145 runs were submitted for scoring. There were 47 runs from 27 groups submitted for the ad hoc task. There were 98 runs submitted from 20 groups for the categorization task. These were distributed across the subtasks of the categorization task as follows: 59 for the triage subtask, 36 for the annotation hierarchy subtask, and three for the annotation hierarchy plus evidence code subtask.

The data from the TREC 2004 Genomics Track are currently available on a password-protected Web site. Information on how to access the data are available at the track Web site.

2 Ad Hoc Retrieval Task

The goal of the ad hoc task was to mimic conventional searching. The use case was a scientist with a specific information need, searching the MEDLINE bibliographic database to find relevant articles to retrieve.

The document collection for the ad hoc retrieval task was a 10-year subset of MEDLINE. We contemplated the use of full-text documents in this task but were unable to procure an adequate amount to represent real-world searching. As such, we chose to use MEDLINE. In reality, despite the widespread availability of on-line, full-text scientific journals at present, most searchers of the biomedical literature still use MEDLINE as an entry point. Consequently, there is great value in being able to search MEDLINE effectively.

The topics for the ad hoc retrieval task were developed from the information needs of real biologists and modified as little as possible to create needs statements with a reasonable

estimated amount of relevant articles (i.e., more than zero but less than one thousand). The information needs capture began with interviews by 12 volunteers who sought biologists in their local environments. A total of 43 interviews yielded 74 information needs. Some of these volunteers, as well as an additional four individuals, created topics in the proposed format from the original interview data. We aimed to have each information need reviewed more than once but were only able to do this with some, ending up with a total of 91 draft topics. The same individuals then were assigned different draft topics for searching on PubMed so they could be modified to generate final topics with a reasonable number of relevant articles. The track chair made one last pass to make the formatting consistent and extract the 50 that seemed most suitable as topics for the track.

Relevance judgments were done using the conventional "pooling method" whereby a fixed number of top-ranking documents from each official run were pooled and provided to an individual (blinded to the number of groups who retrieved the document and what their search statements were). The relevance assessor then judged each document for the specific topic query as definitely relevant (DR), possibly relevant (PR), or not relevant (NR). For the official results, which required binary relevance judgments, documents that were rated DR or PR were considered relevant.

The primary evaluation measure for the task was mean average precision (MAP). Results were calculated using the trec_eval program, a standard scoring system for TREC. A statistical analysis was performed using a repeated measures analysis of variance, with posthoc Tukey tests for pairwise comparisons. In addition to analyzing MAP, we also assessed precision at 10 and 100 documents. The results and analysis of the ad hoc retrieval task are described in more detail in track overview paper.

3 Categorization Task

In the categorization task, we simulated two of the classification activities carried out by human annotators for the MGI system: a triage task and two simplified variations of MGI's annotation task. Systems were required to classify full-text documents from a two-year span (2002-2003) of three journals, with the first year's (2002) documents comprising the training data and the second year's (2003) documents making up the test data.

The documents for the categorization task consisted of articles from three journals over two years, reflecting the full-text documents we were able to obtain from Highwire Press (www.highwire.org). Highwire is a "value added" electronic publisher of scientific journals. Most journals in their collection are published by professional associations, with the copyright remaining with the associations. Highwire originally began with biomedical journals, but in recent years has expanded into other disciplines. They have also supported IR and related research by acting as an intermediary between consenting publishers and information systems research groups who want to use their journals, such as the Genomics Track.

The journals available and used by our track this year were *Journal of Biological Chemistry* (JBC), *Journal of Cell Biology* (JCB), and *Proceedings of the National Academy of Science* (PNAS). These journals have a good proportion of mouse genome articles. Each of the papers from these journals was provided in SGML format based on Highwire's Document Type Definition (DTD). We used articles from the year 2002 for training data and from 2003 for test data. The documents for the categorization tasks came from a subset of articles having the words *mouse, mice* or *murine* as described above. We created a crosswalk file (look-up table) that matched an identifier for each Highwire article (its file name) and its corresponding PubMed ID (PMID). The SGML training document collection was 150 megabytes in size compressed and 449 megabytes uncompressed. The SGML test document collection was 140 megabytes compressed and 397 megabytes uncompressed.

More details about the triage and annotation hierarchy subtasks are available on the track Web site.

4 Future Plans

The TREC Genomics Track will be continuing in 2005. In addition, the data for the 2004 track has been released to the general community for continued experimentation. Planning is currently underway for the tasks and data to be used in the 2005 track.

5 Acknowledgements

The TREC 2004 Genomics Track was supported by NSF Grant ITR-0325160. The track also appreciates the help of Ellen Voorhees and NIST in organizing the track and the National Library of Medicine, Highwire Press, and the Mouse Genome Informatics Project for providing data for use in the track.

6 References

Anonymous (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research,* 32: D258-D261.

Hersh, W., Bhuptiraju, R., et al. (2004). TREC 2004 genomics track overview. *The Thirteenth Text Retrieval Conference: TREC 2004*, Gaithersburg, MD. National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf.

Hirschman, L., Park, J., et al. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics,* 18: 1553-1561.

Mobasheri, A., Airley, R., et al. (2004). Post-genomic applications of tissue microarrays: basic research, prognostic oncology, clinical genomics and drug discovery. *Histology and Histopathology,* 19: 325-335.